VOICE RECOGNITION SYSTEM

BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to a voice recognition system, and specially relates to the speaker adaptive type voice recognition system which is robust to the noise.

2. Description of the Related Art

In the related art, a system shown in Fig. 9 is well known as a speaker adaptive voice recognition system, for example.

This voice recognition system is provided with a previously prepared standard acoustic model 100 of an unspecified speaker, and a speaker adaptive acoustic model 200 is prepared by using a feature vector of an input signal Sc generated from an input voice uttered by a specified speaker, and the standard acoustic model 100, and the voice recognition is conducted by adapting the system to the voice of the specified speaker.

When the adaptive acoustic model 200 is prepared, the standard vector Va corresponding to a designated text (sentence or syllable) Tx is supplied from the standard acoustic model 100 to a path search section 4 and a speaker adaptation section 5, and further, actually, by uttering the designation text Tx by the specified speaker, the input signal Sc is inputted.

Then, after an additive noise reduction section 1 removes

an additive noise included in the input signal Sc, a feature vector generation section 2 generates a feature vector series Vcf which represents the feature quantity of the input signal Sc. Further, a multiplicative noise reduction section 3 removes a multiplicative noise from the feature vector series Vcf, and generates the feature vector series Vc from which the additive noise and the multiplicative noise are removed. The feature vector series Vc is supplied to a path search section 4 and a speaker adaptation section 5.

In this manner, when the standard vector Va and the feature vector series Vc of the input signal Sc actually uttered are supplied to the path search section 4 and the speaker adaptation section 5, the path search section 4 compares the feature vector series Vc to the standard vector Va. Then, the appearance probability of the feature vector series Vc for each syllable, and the state transition probability from an syllable to another syllable are found. Thereafter, when the speaker adaptation section 5 compensates for the standard vector Va according to the appearance probability and the state transition probability, the speaker adaptive acoustic model 200 adaptive to the feature of the voice (input signal) proper to the specified speaker is prepared.

Then, the speaker adaptive acoustic model 200 is adapted to the input signal generated from the uttered voice by the specified speaker. Thereafter, when the specified speaker

utters arbitrarily, the feature vector of the input singal generated from the uttered voice is collated with the adaptive vector of the speaker adaptive acoustic model 200, and the voice recognition is conducted in such a manner that the speaker adaptive acoustic model 200 which gives the highest likelihood is made a recognition result.

In this connection, in the above conventional adaptation type voice recognition system, when the adaptive acoustic model 200 is prepared, the additive noise reduction section 1 removes the additive noise by the spectrum subtraction method, and the multiplicative noise reduction section 3 removes the mulatiplicative noise by the CMN method (cepstrum means normalization), and thereby, the speaker adaptive acoustic model 200 not influenced by the noise is prepared.

That is, the additive noise reduction section 1 removes the spectrum of the additive noise from the spectrum of the input signal Sc after the spectrum of the input signal Sc is found. The mulatiplicative noise reduction section 3 subtracts the time average value from the cepstrum of the input signal Sc after the time average value of the cepstrum of the input signal Sc is found.

However, also in any of the spectrum subtraction method and the CMN method, it is very difficult to remove only noise. Because there is a case where the feature information of the utterance of the speaker proper to be compensated for by the

speaker adaptation is also missed, the adequate speaker adaptive acoustic model 200 cannot be prepared. Therefore, there is a problem that the voice recognition rate is degraded.

SUMMARY OF THE INVENTION

An object of the present invention is to provide a speaker adaptive type voice recognition system which is robust to the noise, to attain the increase of the voice recognition rate.

In order to attain the above object, there is provided a voice recognition system comprising:

a standard acoustic model having a standard vector generated according to information on voice;

a first feature vector generation section for reducing noise from an input signal generated from an uttered voice corresponding to a designated text to generate a first feature vector;

a second feature vector generation section for generating a second feature vector from the input signal having the noise; and

a preparation section for generating an adaptive vector based on the first feature vector, the second feature vector and the standard vector, and preparing a speaker adaptive acoustic model suitable for the uttered voice.

According to the present invention, the preparation section compares the first feature vector with the standard

vector to obtain a path search result; and

the preparation section coordinates the second feature vector with the standard vector according to the path search result to generate the adaptive vector.

According to the present invention, the noise includes additive noise and multiplicative noise.

According to the present invention, the first feature vector generation section includes an additive noise reduction section for reducing the additive noise from the input signal to generate an additive-noise reduced signal.

According to the present invention, the additive noise reduction section applies a transformation to the input signal to generate a first spectrum and subtracting an additive noise spectrum corresponding to the additive noise from the first spectrum.

According to the present invention, the first feature vector generation section includes a cepstrum calculator for applying cepstrum calculation to the additive-noise reduced signal.

According to the present invention, the first feature vector generation section includes a multiplicative noise reduction section for reducing the multiplicative noise by subtracting the multiplicative noise from the first feature vector.

According to the present invention, the first feature

vector contains a plurality of time-series first feature vectors; and

the multiplicative noise reduction section calculates a time average of the time-series first feature vectors for estimating the multiplicative noise.

According to the present invention, the second feature vector generation section applies at least cepstrum calculation to the second spectrum to generate the second feature vector.

According to such the structure, in the case of speaker adaptation, the first feature vector generation section generates the first feature vector except for the additive noise of the peripheral circumstance surrounding the speaker or the multiplicative noise such as transmission noise of the present voice recognition system itself. The second feature vector generation section generates the second feature vector including the additive noise of the peripheral circumstance surrounding the speaker or the feature of the multiplicative noise such as transmission noise of the present voice recognition system itself. Then, the preparation section generates the adaptive vector by compensating the standard vector according to the first feature vector not including the noise and the second feature vector including the noise. Therefore the adoptive vector generates the updated speaker adaptive acoustic model which is adaptive to the voice of the speaker.

As described above, according to the feature vector not

including the noise and the feature vector including the noise, the standard vector in the standard acoustic model is compensated for. Therefore, the speaker adaptive acoustic model corresponding to the practical utterance circumstance can be prepared, and the voice recognition system being robust to the noise and having the higher voice recognition rate can be realized.

Further, the second feature vector generation section generates the feature vector without removing the additive noise or multicative noise and the feature vector is used for the speaker adaptation. Therefore, the feature information of the original voice is not removed, and the adequate speaker adaptive acoustic model can be generated.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram showing the structure of a voice recognition system of an embodiment of the present invention.

Fig. 2 is a table typically showing the structure of a standard acoustic model.

Fig. 3 is a table showing feature vector series $[s_{i,M}]$ generated in a feature vector generation section 12 at the time of speaker adaptation.

Fig. 4 is a table showing feature vector series $[c_{i,M}]$ outputted from a multiplicative noise reduction section 9 at the time of speaker adaptation.

Fig. 5 is a table showing the corresponding relationship of the feature vector series $[c_{i,M}]$ with a standard vector $[a_{n,M}]$ according to the frame number and the state number.

Fig. 6 is a table showing the relationship of the feature vector series $[c_{i,M}]$, the standard vector $[a_{n,M}]$, the frame number and the state number.

Fig. 7 is a table showing the relationship of the average feature vector generated by the speaker adaptation with the standard vector.

Fig. 8 is a table showing the content of the speaker adaptive acoustic model after update.

Fig. 9 is a block diagram showing the structure of a speaker adaptation type voice recognition system in the related art.

DETAILED DESCRIPTION OF THE PRESENT INVENTION

Referring to the drawings, the present invention will be described below with reference to the accompanying drawings. In this connection, Fig. 1 is a block diagram showing the structure of a voice recognition system according to an embodiment of the present invention.

In Fig. 1, the voice recognition system comprises the standard acoustic model (hereinafter, referred to as [standard voice HMM]) 300 of an unspecified speaker previously prepared by using the Hidden Markov model (HMM) and the speaker adaptation acoustic model (hereinafter, referred to as [adaptive voice

HMM]) 400 prepared by the speaker adaptation.

In this connection, to easily understand the embodiment of the present invention, the state number of the standard voice HMM 300 is defined as 1. Further, the standard voice HMM 300 has an appearance probability distribution for each syllable, and an average vector of the appearance probability distribution is to be a standard vector.

Accordingly, as typically shown in Fig. 2, the standard voice HMM 300 has a M dimensional standard vector $[a_{n,M}]$ for each syllable. That is, when the standard voice HMM 300 is prepared, for example, voice data generated from an uttered voice by one or plurality of speakers (unspecified speakers) under silent environment is framed for each predetermine time. The framed voice data is successively cepstrum-operated, to generate the feature vector series in the cepstrum domain for a plurality of frames for each syllable. Obtaining the average of the feature vector series for a plurality of frames prepares the standard voice HMM 300 composed of the standard vector $[a_{n,M}]$ for each syllable.

Herein, a variable n of the standard vector $[a_{n,M}]$ expresses the state number to recognize each syllable, and a variable M expresses the dimension of the vector. For example, the Japanese syllable [A] corresponding to the state number $n = 1$ is characterized as the M dimensional standard vector $[a_{n,M}] = [a_{1,1}, a_{1,2}, a_{1,3}, \ldots a_{1,M}]$, and the Japanese syllable

[I] corresponding to the state number n = 2 is characterized as the M dimensional standard vector $[a_{n,M}] = [a_{2,1}, a_{2,2}, a_{2,3}, \ldots a_{2,M}]$. The same rule applies correspondingly to the following, and the remaining syllables are also characterized as the M dimensional standard vector $[a_{n,M}]$ distinguished by the state number n.

At the time of the speaker adaptation which will be described later, the standard voice HMM 300 is supplied with the designated text Tx of the previously determined sentence or syllable, and the standard vector $[a_{n,M}]$ corresponding to the syllable structuring the designated text Tx is supplied to the path search section 10 and the speaker adaptation section 11, according to the arrangement sequence of the syllable.

For example, when the designated text Tx of Japanese [KONNICHIWA] is supplied, the standard vectors corresponding to respective state numbers n = 10, 46, 22, 17, 44 expressing [KO], [N], [NI], [CHI], [WA], $[a_{10,1}, a_{10,2}, a_{10,3}, \ldots a_{10,M}]$, $[a_{46,1}, a_{46,2}, a_{46,3}, \ldots a_{46,M}]$, $[a_{22,1}, a_{22,2}, a_{22,3}, \ldots a_{22,M}]$, $[a_{17,1}, a_{17,2}, a_{17,3}, \ldots a_{17,M}]$, and $[a_{44,1}, a_{44,2}, a_{44,3}, \ldots a_{44,M}]$ are supplied to the path search section 10 and the speaker adaptation section 11 in order.

Further, the voice recognition system of the present invention is provided with a framing section 6, additive noise reduction section 7, feature vector generation section 8, multiplicative noise reduction section 9, and feature

vector generation section 12.

The framing section 6, when the specified speaker actually utters the designated text Tx at the time of the speaker adaptation, divides an input signal Sc into generated from the uttered voice frames for each predetermined time (for example, 10 - 20 msec) and outputs it to the additive noise reduction sections 7, 13 and feature vector generation section 12.

The additive noise reduction section 7 successively conducts Fourier transformation on each framed input signal Scf divided into each frame to generate a spectrum for each frame. Further, the additive noise included in each spectrum is removed in the spectrum domain to output the spectrum.

The feature vector generation section 8 conducts the cepstrum operation on the spectrum having no additive noise for each frame to generate the feature vector series $[c_i, _M]'$ in the cepstrum domain. In this connection, the variable i of the feature vector series $[c_i, _M]'$ expresses the order (number), and the variable M expresses the dimension.

The multiplicative noise reduction section 9 removes the multiplicative noise from the feature vector series $[c_i, _M]'$ by using the CMN method. That is, a plurality of the feature vector series $[c_i, _M]'$ obtained for each frame i by the feature vector generation section 8 are time-averaged for each dimension. When the M dimensional time average value $[c\hat{}_M]$ obtained thereby is subtracted from each feature vector $[c_i, _M]'$ to generate the

feature vector series $[c_{i,M}]$ from which the maultiplicative noise is removed. The feature vector series $[c_{i,M}]$ thus generated is supplied to the path search section 10.

The feature vector generation section 12 generates spectrum for the frame when each framed input signal Scf divided for frame outputted from the framing section 6 is successively Fourier-transformed. Further, when each spectrum is cepstrum-operated for each frame, the feature vector series $[s_{i,M}]$ in the cepstrum domain is generated and supplied to the speaker adaptation section 11. In this connection, the variable i of the feature vector series $[s_{i,M}]$ expresses the order for each frame, and the variable M expresses the dimension.

In this manner, the designation text Tx, standard vector $[a_{n,M}]$ and feature vector series $[c_{i,M}]$ are supplied to the path search section 10. The designation text Tx, standard vector $[a_{n,M}]$ and feature vector series $[s_{i,M}]$ are supplied to the speaker adaptation section 11.

The path search section 10 compares the standard vector $[a_{n,M}]$ to the feature vector series $[c_{i,M}]$, and judges which syllable of the designation text Tx corresponds to the feature vector series $[c_{i,M}]$ for each frame. The path search result Dv is supplied to the speaker adaptation section 11.

The speaker adaptation section 11 divides the feature vector series $[s_{i,M}]$ from the feature vector generation section 12 into each syllable according to the path search result Dv.

Then, the average is obtained for each dimension with respect to the feature vector series $[s_{i,M}]$ for each divided syllable. Eventually, the average feature vector $[s\hat{}_{n,M}]$ for each syllable is generated.

Further, the speaker adaptation section 11 finds a difference vector $[d_{n,M}]$ between the standard vector $[a_{n,M}]$ of each syllable corresponding to the designation text Tx, and the average feature vector $[s\hat{}_{n,M}]$. Then, conducting the average operation on these difference vectors $[d_{n,M}]$ leads to finding the M dimensional movement vector $[m_M]$ expressing the feature of the specified speaker. Further, the adaptation vectors $[x_{n,M}]$ for all syllables are generated by adding the movement vector $[m_M]$ to the standard vectors $[a_{n,M}]$ of all syllables from the standard voice HMM 300. The adaptive voice HMM 400 is updated by these adaptation vectors $[x_{n,M}]$.

Next, referring to Fig. 2 - Fig. 8, the function of the path search section 10 and the speaker adaptation section 11 will be described in detail.

In this connection, the designation text Tx of Japanese [KONNICHIWA] is used as a typical example.

Further, it is defined that the input signal Sc of Japanese [KONNICHIWA] uttered by the speaker is divided into 30 frames by the framing section 6 and inputted.

The standard voice HMM 300 is, as shown in Fig. 2, prepared as the standard vector $[a_{n,M}]$ of the unspecified speaker

corresponding to each of a plurality of syllables. Further, each syllable is classified by the state number n.

Further, the adaptive voice HMM 400 is set to the same content (default setting) as the standard vector $[a_{n,M}]$ of the standard voice HMM 300 before the speaker adaptation, as shown in Fig. 2.

At the beginning of the speaker adaptation processing, the designation text Tx of Japanese [KONNICHIWA] is supplied to the standard voice HMM 300. Then, supplied to the path search section 10 and the speaker adaptation section 11 are the standard vector $[a_{10,1}, a_{10,2}, a_{10,3}, \ldots a_{10,M}]$ corresponding to the state number n = 10 expressing the syllable [KO], the standard vector $[a_{46,1}, a_{46,2}, a_{46,3}, \ldots a_{46,M}]$ corresponding to the state number n = 46 expressing the syllable [N], the standard vector $[a_{22,1}, a_{22,2}, a_{22,3}, \ldots a_{22,M}]$ corresponding to the state number n = 22 expressing the syllable [NI], the standard vector $[a_{17,1}, a_{17,2}, a_{17,3}, \ldots a_{17,M}]$ corresponding to the state number n = 17 expressing the syllable [CHI], and the standard vector $[a_{44,1}, a_{44,2}, a_{44,3}, \ldots a_{44,M}]$ corresponding to the state number n = 44 expressing the syllable [WA].

Next, when the speaker utters [KONNICHIWA], the framing section 6 divides the input signal Sc into 30 frames according to the lapse of time, and outputs the divided input signal Sc. Then, the feature vector generation section 12 generates the feature vectors $[s_{1,1}, s_{1,2}, s_{1,3}, \ldots s_{1,M}] - [s_{30,1}, s_{30,2}, s_{30},$

$_{3,}$ ... $s_{30, M}$] of the framed input signal Scf according to the order of each frame, and supplies to the speaker adaptation section 11.

That is, as typically shown in Fig. 3, the feature vector generation section 12 generates the feature vector series for 30 frames of i = 1 - 30, $[s_{i, M}] = [s_{1, 1}, s_{1, 2}, s_{1, 3}, \ldots s_{1, M}] - [s_{30, 1}, s_{30, 2}, s_{30, 3}, \ldots s_{30, M}]$, and supplies to the speaker adaptation section 11.

On the one hand, the processing system includes the additive noise reduction section 7, feature vector generation section 8, and multiplicative noise reduction section 9. In the processing system, the feature vector series $[c_{1, M}] = [c_{1, 1}, c_{1, 2}, c_{1, 3}, \ldots c_{1, M}] - [c_{30, 1}, c_{30, 2}, c_{30, 3}, \ldots c_{30, M}]$ for 30 frames of i = 1 - 30 are generated according to the framed input signal Scf of each frame supplied from the framing section 6, and supplied to the path search section 10. That is, as typically shown in Fig. 4, the feature vector series for 30 frames $[c_{i, M}] = [c_{1, 1}, c_{1, 2}, c_{1, 3}, \ldots c_{1, M}] - [c_{30, 1}, c_{30, 2}, c_{30, 3}, \ldots c_{30, M}]$ are supplied to the path search section 10 through the multiplicative noise reduction section 9.

The path search section 10 compares the feature vector series $[c_{1, M}]$ for 30 frames to the standard vector $[a_{n, M}]$ corresponding to each syllable of the designation text Tx, by the methods of Viterbi algorithm or forward backward algorithm, and finds which syllable corresponds to the feature vector series

$[c_{1, M}]$ at each moment for each frame.

Thereby, as shown in Fig. 5, each frame number i of 30 frames is coordinated to each state number n expressing each syllable of [KONNICHIWA]. Then, the coordinated result is supplied to the speaker adaptation section 11 as the path search result Dv.

The speaker adaptation section 11 coordinates the feature vectors $[s_{1, 1}, s_{1, 2}, s_{1, 3}, \ldots s_{1, M}] - [s_{30, 1}, s_{30, 2}, s_{30, 3}, \ldots s_{30, M}]$ to the standard vectors $[a_{10, 1}, a_{10, 2}, a_{10, 3}, \ldots a_{10, M}]$, $[a_{46, 1}, a_{46, 2}, a_{46, 3}, \ldots a_{46, M}]$, $[a_{22, 1}, a_{22, 2}, a_{22, 3}, \ldots a_{22, M}]$, $[a_{17, 1}, a_{17, 2}, a_{17, 3}, \ldots a_{17, M}]$, $[a_{44, 1}, a_{44, 2}, a_{44, 3}, \ldots a_{44, M}]$, according to the path search result Dv.

That is, as shown in Fig. 6, the standard vector $[a_{10, 1}, a_{10, 2}, a_{10, 3}, \ldots a_{10, M}]$ is coordinated to the feature vector $[s_{1, 1}, s_{1, 2}, s_{1, 3}, \ldots s_{1, M}] - [s_{6, 1}, s_{6, 2}, s_{6, 3}, \ldots s_{6, M}]$ of the frame number i = 1 - 6 corresponding to the syllable [KO] obtained by the path search. The standard vector $[a_{46, 1}, a_{46, 2}, a_{46, 3}, \ldots a_{46, M}]$ is coordinated to the feature vector $[s_{7, 1}, s_{7, 2}, s_{7, 3}, \ldots s_{7, M}] - [s_{10, 1}, s_{10, 2}, s_{10, 3}, \ldots s_{10, M}]$ of the frame number i = 7 - 10 corresponding to the syllable [N].

Further, the standard vector $[a_{22, 1}, a_{22, 2}, a_{22, 3}, \ldots a_{22, M}]$ is coordinated to the feature vector $[s_{11, 1}, s_{11, 2}, s_{11, 3}, \ldots s_{11, M}] - [s_{14, 1}, s_{14, 2}, s_{14, 3}, \ldots s_{14, M}]$ of the frame number i = 11 - 14 corresponding to the syllable [NI]. The standard vector $[a_{17, 1}, a_{17, 2}, a_{17, 3}, \ldots a_{17, M}]$ is coordinated to the

feature vector $[S_{15,1}, S_{15,2}, S_{15,3}, \ldots S_{15,M}] - [S_{18,1}, S_{18,2}, S_{18,3}, \ldots S_{18,M}]$ of the frame number i = 15 - 18 corresponding to the syllable [CHI]. The standard vector $[a_{44,1}, a_{44,2}, a_{44,3}, \ldots a_{44,M}]$ is coordinated to the feature vector $[S_{19,1}, S_{19,2}, S_{19,3}, \ldots S_{19,M}] - [S_{30,1}, S_{30,2}, S_{30,3}, \ldots S_{30,M}]$ of the frame number i = 19 - 30 corresponding to the syllable [WA].

Next, the speaker adaptation section 11 divides the feature vectors $[S_{1,1}, S_{1,2}, S_{1,3}, \ldots S_{1,M}] - [S_{30,1}, S_{30,2}, S_{30,3}, \ldots S_{30,M}]$ for 30 frames shown in Fig. 6, for each syllable of [KO], [N], [NI], [CHI], [WA]. As shown in Fig. 7, the average feature vector for each syllable of [KO], [N], [NI], [CHI], [WA], $[s\hat{}_{n,M}]$ is generated by obtaining the average for each divided feature vector.

That is, relating to the feature vectors $[S_{1,1}, S_{1,2}, S_{1,3}, \ldots, S_{1,M}] - [S_{6,1}, S_{6,2}, S_{6,3}, \ldots, S_{6,M}]$ of the first - sixth frames (frame number k = 6) corresponding to the syllable [KO] shown in Fig. 6, as shown by the following expression (1), the first dimensional 6 elements $S_{1,1} - S_{6,1}$ are added, and the first dimensional element $s\hat{}_{n,1}$ of the average feature vector $[s\hat{}_{n,M}]$ is obtained by multiplying the added value $(S_{1,1} + s_{2,1} + S_{3,1} + S_{4,1} + S_{5,1} + S_{6,1})$ by frame number k = 6. Further, in the same manner as to the second dimensional 6 elements $S_{1,2} - S_{6,2}$, the added value $(S_{1,2} + S_{2,2} + S_{3,2} + S_{4,2} + S_{5,2} + S_{6,2})$ is obtained. Then, the second dimensional element $s\hat{}_{n,2}$ of the average feature vector $[s\hat{}_{n,M}]$ is obtained

by multiplying it by frame number k = 6. In the same manner as also in the following, the element $s\hat{}_{n,M}$ up to the M dimensional 6 element $s_{1,M} - s_{6,M}$ is obtained, and the M dimensional average feature vector [$s\hat{}_{n,1}$, $s\hat{}_{n,2}$, $s\hat{}_{n,3}$, ... $s\hat{}_{n,M}$] corresponding to the syllable [KO] composed of these M dimensional elements $s\hat{}_{n,1} - s\hat{}_{n,M}$ is generated.

$$s\hat{}_{n,M} = (s_{1,1} + s_{2,1} + s_{3,1} + s_{4,1} + s_{5,1} + s_{6,1})/k \quad \ldots (1)$$

where the variable k in the expression (1) is the frame number in each syllable;

the variable n is the state number to distinguish each syllable; and

the variable M expresses the dimension.

Accordingly, the variable n in the expression (1) is n = 10, and the M dimensional average feature vector corresponding to the syllable [KO] is [$s\hat{}_{10,1}$, $s\hat{}_{10,2}$, $s\hat{}_{10,3}$, ... $s\hat{}_{10,M}$].

Further, the average feature vector [$s\hat{}_{46,1}$, ... $s\hat{}_{46,M}$] corresponding to the remaining syllable [N], the average feature vector [$s\hat{}_{22,1}$, ... $s\hat{}_{22,M}$] corresponding to the syllable [NI], the average feature vector [$s\hat{}_{17,1}$, ... $s\hat{}_{17,M}$] corresponding to the syllable [CHI], and the average feature vector [$s\hat{}_{44,1}$, ... $s\hat{}_{44,M}$] corresponding to the syllable

[WA], are also obtained in the same manner.

Next, according to the next expression (2), the difference vector $[d_{10,1}, \ldots d_{10,M}]$, $[d_{46,1}, \ldots d_{46,M}]$, $[d_{22,1}, \ldots d_{22,M}]$, $[d_{17,1}, \ldots d_{17,M}]$, and $[d_{44,1}, \ldots d_{44,M}]$ between the average feature vector $[s\hat{}_{10,1}, \ldots s\hat{}_{10,M}]$, $[s\hat{}_{46,1}, \ldots s\hat{}_{46,M}]$, $[s\hat{}_{22,1}, \ldots s\hat{}_{22,M}]$, $[s\hat{}_{17,1}, \ldots s\hat{}_{17,M}]$, $[s\hat{}_{44,1} \ldots s\hat{}_{44,M}]$ corresponding to each of syllables [KO], [N], [NI], [CHI], and [WA], and the standard vector $[a_{10,1}, \ldots a_{10,M}]$, $[a_{46,1}, \ldots a_{46,M}]$, $[a_{22,1}, \ldots a_{22,M}]$, $[a_{17,1}, \ldots a_{17,M}]$, and $[a_{44,1}, \ldots a_{44,M}]$, are respectively obtained.

$$d_{n,j} = s\hat{}_{n,j} - a_{n,j} \qquad \ldots (2)$$

where the variable n in the expression (2) shows the state numbers n = 10, 46, 22, 17, 44, corresponding to each of syllables [KO], [N], [NI], [CHI], [WA]; and

the variable j shows each of dimensions j = 1 - M of the vector.

Then, the obtained difference vectors $[d_{10,1}, \ldots d_{10,M}]$, $[d_{46,1}, \ldots d_{46,M}]$, $[d_{22,1}, \ldots d_{22,M}]$, $[d_{17,1}, \ldots d_{17,M}]$, and $[d_{44,1}, \ldots d_{44,M}]$ are applied to the next expression (3). The M dimensional movement vector $[m_M] = [m_1, m_2, \ldots, m_M]$ of these 5 (V = 5) syllables of [KO], [N], [NI], [CHI], [WA] is obtained from the average for each dimension.

$$m_j = \frac{1}{V}\sum_n d_{n,j} \quad \ldots \quad (3)$$

where the variable j in the expression (3) shows each dimension j = 1 - M of the vector;

the variable n shows the state numbers n = 10, 46, 22, 17, 44 corresponding to each of syllables [KO], [N], [NI], [CHI], [WA]; and

the variable V shows the number of syllables (V = 5).

Thus obtained movement vector $[m_M] = [m_1, m_2, \ldots, m_M]$ expresses the feature of the specified speaker. Then, as shown by the next operational expression (4), the adaptive vector $[x_{n,M}]$ having the feature proper to the speaker is obtained from addition of the movement vector $[m_M]$ to the standard vector $[a_{n,M}]$ of the all syllables. Further, as shown in Fig. 8, the processing of the speaker adaptation is completed by updating the adaptive voice HMM 400 by the obtained adaptive vector $[x_{n,M}]$.

$$[x_{n,M}] = [a_{n,M}] + [m_M] \quad \ldots \quad (4)$$

It is described hereinabove that the adaptive voice HMM 400 has the speaker adaptation according to the designation text Tx of [KONNICHIWA]. However, when the adaptive voice HMM 400 has the speaker adaptation according to the designation

text Tx including other syllables, all the syllable in the adaptive voice HMM 400 can also have the speaker adaptation.

Next, after the speaker adaptation generates the adaptive voice HMM 400, when the specified speaker conducts an arbitrary utterance, the framing section 6 divides the input signal Sc into the frames for each predetermined time (for example, 10 - 20 msec) in the same manner as the above. Then, the framing section 6 outputs the framed input signal Scf of each frame according to the lapse of time, and supplies to the additive noise reduction section 13.

The additive noise reduction section 13, in the same manner as the above additive noise reduction section 7, conducts Fourier transformation on each framed input signal Scf divided for frame, and generates the spectrum for each frame. Further, the additive noise reduction section 13 removes the additive noise included in each spectrum in the spectrum domain, and outputs the spectrum to the feature vector generation section 14.

The feature vector generation section 14, in the same manner as in the above feature vector generation section 8, conducts the cepstrum operation on the spectrum having no additive noise for frame, generates the feature vector series $[y_{i,M}]'$ in the cepstrum domain, and outputs to the multiplicative noise reduction section 15.

The multiplicative noise reduction section 15, in the same manner as in the above multiplicative noise reduction

section 9, removes the multiplicative noise from the feature vector series $[y_{i,M}]'$ by using the CMN method, and supplies the M dimensional feature vector series $[y_{i,M}]$ having no multiplicative noise, to the recognition section 16. Here, the variable i of the feature vector series $[y_{i,M}]$ expresses the frame number.

As described above, when the feature vector series $[y_{i,M}]$ based on the input signal generated from the actually uttered voice is supplied to the recognition section 16, the recognition section 16 collates the feature vector series $[y_{i,M}]$ with the adaptive vector $[x_{n,M}]$ of the adaptive voice HMM 400 in which the speaker adaptation is conducted, and outputs the adaptive voice HMM 400 which gives the highest likelihood as the recognition result.

As described above, according to the voice recognition system of the present invention, when the specified speaker utters the designation text Tx upon speaker adaptation, the additive noise reduction section 7, feature vector generation section 8 and multiplicative noise reduction section 9 generate the feature vector series $[c_{i,M}]$ from which the additive noise and multipicative noise are removed. The feature vector generation section 12 generates the feature vector series $[s_{i,M}]$ according to the framed input signal Scf including the additive noise and multipicative noise. The path search section 10 and speaker adaptation section 11 generate the adaptive vector $[x_i,$

$_{M}$] according to these feature vector series [$c_{i,M}$], feature vector series [$s_{i,M}$], and standard vector [$a_{i,M}$]. The adaptive vector [$x_{i,M}$] in which the speaker adaptation is conducted updates the adaptive voice HMM 400.

Accordingly, the feature vector series [si, M] including the feature of the noise (additive noise) of the peripheral circumstance surrounding the specified speaker, or transmission noise (multiplicative noise) of the present voice recognition system itself, is used for the speaker adaptation. Therefore, the adaptive voice HMM 400 corresponding to the actual utterance circumstance can be generated from the voice recognition system which is robust to the noise and whose voice recognition rate is high.

Further, in the speaker adaptation type voice recognition system in the related art, at the time of the speaker adaptation, the generation of the feature vector from which the additive noise and multiplicative noise are removed misses the feature information of the utterance proper to the speaker to be compensated for by the speaker adaptation. There is a problem that the adequate speaker adaptive acoustic model cannot be prepared.

On the other hand, according to the voice recognition system of the present invention, the feature vector generation section 12 generates the feature vector series [$s_{i,M}$] without removing the additive noise and multiplicative noise. The

feature information of the utterance proper to the speaker to be compensated for by the speaker adaptation is not missed because the feature vector series $[s_{i,M}]$ is used for the speaker adaptation. Therefore, the adequate speaker adaptive acoustic model can be prepared to increase the voice recognition rate.

In this connection, in the present invention, it is described that the adaptive voice HMM 400 on the basis of the syllables such as Japanese [AIUEO] is prepared. However, it is not limited to only the syllable, but, the adaptive voice HMM 400 based on the phoneme can be prepared.

Further, in the present invention, a simple example is taken as an example, and the method of the speaker adaptation is described. However, the speaker adaptation method of the present invention can be adapted for other various speaker adaptation methods in which the standard vector $[a_{n,M}]$ is coordinated with the feature vector series $[s_{i,M}]$ or $[c_{i,M}]$ of the speaker adaptation. According thereto, the speaker adaptive acoustic model can be generated.

As described above, according to the voice recognition system of the present invention, when the speaker adaptation is conducted, the feature vector from which the additive noise or the multiplicative noise is removed, and the feature vector including the feature of the additive noise or the multiplicative noise are generated. According to the feature vector not

including noise and the feature vector including the noise, the standard vector is compensated for. Because the speaker adaptive acoustic model adaptive for the utterance proper to the speaker is prepared, the speaker adaptive acoustic model adaptive for the actual utterance circumstance can be generated.

Further, because the feature vector is used for the speaker adaptation without removing the additive noise or multiplicative noise, the feature information of the utterance proper to the speaker to be compensated for by the speaker adaptation, is not missed. Therefore, an adequate speaker adaptive acoustic model can be generated.

Therefore, a voice recognition system being robust to the noise and whose voice recognition rate is high can be obtained.